DEGREE DISTRIBUTION IN THE LOWER LEVELS OF THE UNIFORM RECURSIVE TREE

ÁGNES BACKHAUSZ AND TAMÁS F. MÓRI

Department of Probability Theory and Statistics, Faculty of Science
Eötvös Loránd University
Pázmány P. s. 1/C, H-1117 Budapest, Hungary
E-mail address: `agnes@cs.elte.hu, moritamas@ludens.elte.hu`

ABSTRACT. In this note we consider the kth level of the uniform random recursive tree after n steps, and prove that the proportion of nodes with degree greater than $t \log n$ converges to $(1 - t)^k$ almost surely, as $n \to \infty$, for every $t \in (0, 1)$. In addition, we show that the number of degree d nodes in the first level is asymptotically Poisson distributed with mean 1; moreover, they are asymptotically independent for $d = 1, 2, \ldots$.

1. INTRODUCTION

Let us consider the following random graph model. We start from a single node labelled with 0. At the nth step we choose a vertex at random, with equal probability, and independently of the past. Then a new node, vertex n, is added to the graph, and it is connected to the chosen vertex. In this way a random tree, the so called uniform recursive tree, is built.

This model has a long and rich history. Apparently, the first publication where the uniform recursive tree appeared was [11]. Since then a huge number of papers have explored the properties of this simple combinatorial structure.

Recursive trees serve as probabilistic models for system generation, spread of contamination of organisms, pyramid scheme, stemma construction of philology, Internet interface map, stochastic growth of networks, and many other areas of application, see [6] for references. For a survey of probabilistic properties of uniform recursive trees see [5] or [9]. Among others, it is known that this random tree has an asymptotic degree distribution, namely, the proportion of nodes with degree d converges, as $n \to \infty$, to 2^{-d} almost surely. Another important quantity is the maximal degree, which is known to be asymptotically equal to $\log_2 n$ [4]. Considering our graph a rooted tree, we can define the levels of the tree in the usual way: level k is the set $L_n(k)$ of the vertices that are of distance k from vertex 0, the root. It is not hard to find the a.s. asymptotics of the size of level k after step n; it is

$$|L_n(k)| \sim \mathbb{E}|L_n(k)| \sim \frac{(\log n)^k}{k!}, \quad k = 1, 2, \ldots.$$

Recursive trees on nodes $0, 1, \ldots, n-1$ can be transformed into permutations $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ in the following recursive way. Start from the identity permutation $\sigma = (1, 2, \ldots, n)$. Then, taking the nodes $1, 2, \ldots, n-1$ one after another, update the permutation by swapping σ_{i+1} and σ_{i+1-j} if node i was connected to node $j < i$ at the time it was added to the tree. It is easy to see that in this way a one-to-one correspondence is set between trees and permutations, and the uniform recursive tree is transformed into a uniform random permutation.

Another popular recursive tree model is the so called plane oriented recursive tree. It was originally proposed by Szymański [10], but it got in the focus of research after the seminal paper of Barabási and Albert [3]. A non-oriented version of it starts from a single edge, and at each step a new vertex is added to the graph. The new vertex is then connected to one of the old nodes at random; the other endpoint of the new edge is chosen from the existing vertices with probability proportional to the instanteneous degree of the node (preferential attachment). This can also be done in such a way that we select an edge at random with equal probability, then choose one of its endpoints. In this tree the proportion of degree d nodes converges to $\frac{4}{d(d+1)(d+2)}$ with probability 1.

Katona has shown [7] that the same degree distribution can be observed if one is confined to any of the largest levels. On the other hand, if we only consider a fixed level, the asymptotic degree distribution still exists, but it becomes different [8]. This phenomenon has been observed in other random graphs, too. A general result of that kind has been published recently [2].

In the present note we will investigate the lower levels of the uniform recursive tree. We will show that, unlike in many scale free recursive tree models, no asymptotic degree distribution emerges. Instead, for almost all nodes in the lower levels the degree sequence grows to infinity at the same rate as the overall maximum of degrees does. We also investigate the number of degree d vertices in the first level for $d = 1, 2, \ldots$, and show that they are asymptotically i.i.d. Poisson with mean 1.

2. Nodes of high degree in the lower levels

Let $\deg_n(i)$ denote the degree of node i after step n ($i \le n$). Further, let $Z_{n,k}(t)$ denote the proportion of nodes in level k with degree greater than $t \log n$. Formally,

$$Z_{n,k}(t) = \frac{1}{|L_n(k)|} \left| \left\{ i \le n : i \in L_n(k), \ \deg_n(i) > t \log n \right\} \right|.$$

The main result of this section is the following theorem.

Theorem 2.1. *For $k = 1, 2, \ldots$ and $0 < t < 1$*

$$\lim_{n \to \infty} Z_{n,k}(t) = (1-t)^k \quad a.s.$$

For the proof we need some auxiliary lemmas, interesting in their own right.

Let the number n of steps be fixed, and $1 < i < n$. Firstly, we are interested in $X = \deg_n(i) - 1$.

Lemma 2.1. *Let $0 < \varepsilon < t < 1$. Then for every $i > n^{1-t+\varepsilon}$ we have*

$$\mathbb{P}(X > t \log n) \le \exp\left(-\frac{\varepsilon^2}{2t} \log n \right).$$

Proof. $X = I_{i+1} + I_{i+2} + \cdots + I_n$, where $I_j = 1$, if vertex i gets a new edge at step n, and 0 otherwise. These indicators are clearly independent and $EI_j = 1/j$, hence

$$\mathbb{E}X = \frac{1}{i+1} + \cdots + \frac{1}{n}.$$

Let us abbreviate it by s. Clearly,

$$\log \frac{n}{i+1} \le s \le \log \frac{n}{i}.$$

Let $a > s$, then by [1, Theorem A.1.12] we have

$$\mathbb{P}(X \ge a) \le \left(e^{\beta - 1} \beta^{-\beta}\right)^s,$$

where $\beta = a/s$. Hence

$$\mathbb{P}(X \ge a) \le e^{a-s} \left(\frac{s}{a}\right)^a = e^{a-s}\left(1 - \frac{a-s}{a}\right)^a$$

$$= \exp\left(a - s - a\left(\frac{a-s}{a} + \frac{1}{2}\left(\frac{a-s}{a}\right)^2 + \dots\right)\right) \le \exp\left(-\frac{(a-s)^2}{2a}\right).$$

Now, set $a = t \log n$. Then $s \le (t - \varepsilon) \log n$, and

$$\mathbb{P}(X \ge t \log n) \le \exp\left(-\frac{(t \log n - s)^2}{2t \log n}\right) \le \exp\left(-\frac{\varepsilon^2}{2t} \log n\right).$$

$\square$

Lemma 2.2. *Let $0 < t < 1$, and $0 < \varepsilon < 1 - t$. Then for every $i \le n^{1-t-\varepsilon} - 1$ we have*

$$\mathbb{P}(X \le t \log n) \le \exp\left(-\frac{\varepsilon^2}{2(t+\varepsilon)} \log n\right).$$

Proof. This time $s > \log \frac{n}{i+1} \ge (t + \varepsilon) \log n$, thus [1, Theorem A.1.13] implies that

$$\mathbb{P}(X \le t \log n) \le \exp\left(-\frac{(s - t \log n)^2}{2s}\right).$$

Notice that the exponent in the right-hand side, as a function of s, is decreasing for $s > t \log n$. Therefore s can be replaced by $(t + \varepsilon) \log n$, and the proof is complete. $\square$

Proof of Theorem 2.1. Since $\deg_n(i)$ is approximately equal to $\log \frac{n}{i}$, it follows that $\deg_n(i) \ge t \log n$ is approximately equivalent to $i \le n^{1-t}$. Basing on Lemmas 2.1 and 2.2 we can quantify this heuristic reasoning.

Let $0 < \varepsilon < \min\{t, 1 - t\}$, and $a = a(n) = \lfloor n^{1-t-\varepsilon} \rfloor - 1$, $b = b(n) = \lceil n^{1-t+\varepsilon} \rceil$. Then by Lemma 2.2

$$\mathbb{P}\big(\exists i \in L_n(k) \text{ such that } i \le a, \ \deg_n(i) \le 1 + t \log n\big)$$

$$\le \sum_{i=1}^{a} \mathbb{P}\big(i \in L_n(k), \ \deg_n(i) \le 1 + t \log n\big)$$

$$= \sum_{i=1}^{a} \mathbb{P}\big(i \in L_n(k)\big)\mathbb{P}\big(\deg_n(i) \le 1 + t \log n\big)$$

$$\le \mathbb{E}L_n(k) \cdot \exp\left(-\frac{\varepsilon^2}{2(t+\varepsilon)} \log n\right).$$

Similarly, by Lemma 2.1,

$$\mathbb{P}\big(\exists i \in L_n(k) \text{ such that } i > b,\ \deg_n(i) > 1 + t \log n\big)$$

$$\leq \sum_{i=b+1}^{n} \mathbb{P}\big(i \in L_n(k),\ \deg_n(i) > 1 + t \log n\big)$$

$$= \sum_{i=b+1}^{n} \mathbb{P}\big(i \in L_n(k)\big)\mathbb{P}\big(\deg_n(i) > 1 + t \log n\big)$$

$$\leq \mathbb{E}L_n(k) \cdot \exp\left(-\frac{\varepsilon^2}{2t} \log n\right).$$

Introduce the events

$$A(n) = \big\{L_a(k) \subset \{i \in L_n(k) : \deg_n(i) > 1 + t \log n\} \subset L_b(k)\big\},$$

then the probability of their complements can be estimated as follows.

$$\mathbb{P}\left(\overline{A(n)}\right) \leq 2\mathbb{E}|L_n(k)| \exp\left(-\frac{\varepsilon^2}{2t} \log n\right).$$

Note that $|L_a(k)| \sim (1 - t - \varepsilon)^k |L_n(k)|$, and $|L_b(k)| \sim (1 - t + \varepsilon)^k |L_n(k)|$, a.s.

Let $c > 2(t + \varepsilon)\varepsilon^{-2}$, then $\sum_{n=1}^{\infty} \mathbb{P}\left(\overline{A(n^c)}\right) < \infty$, hence by the Borel–Cantelli lemma it follows almost surely that $A(n^c)$ occurs for every n large enough. Consequently,

$$(1 - t - \varepsilon)^k |L_{n^c}(k)| \big(1 + o(1)\big)$$

$$\leq |\{i \in L_{n^c}(k) : \deg_{n^c}(i) > 1 + t \log(n^c)\}|$$

$$\leq (1 - t + \varepsilon)^k |L_{n^c}(k)| \big(1 + o(1)\big).$$

This implies

$$\liminf_{n\to\infty} Z_{n^c,k}(t) \geq (1 - t - \varepsilon)^k \text{ and } \limsup_{n\to\infty} Z_{n^c,k}(t) \leq (1 - t + \varepsilon)^k$$

for every positive ε, hence Theorem 2.1 is proven along the subsequence (n^c).

To the indices in between we can apply the following esimation. For $n^c \leq N \leq (n+1)^c$ with sufficiently large n we have

$$Z_{N,k}(t) \leq \frac{1}{|L_{n^c}(k)|} \left|\left\{i \in L_{(n+1)^c}(k) : \deg_{(n+1)^c}(i) \geq t \log(n^c)\right\}\right|$$

$$= \frac{|L_{(n+1)^c}(k)|}{|L_{n^c}(k)|} Z_{(n+1)^c,k}\left(t \frac{\log n}{\log(n+1)}\right).$$

Here the first term tends to 1, while the second term's asymptotic behaviour is just the same as that of $Z_{(n+1)^c,k}(t)$. Hence $Z_{N,k}(t) \leq \big(1 + o(1)\big)(1 - t)^k$.

Similarly,

$$Z_{N,k}(t) \geq \frac{|L_{n^c}(k)|}{|L_{(n+1)^c}(k)|} Z_{n^c,k}\left(t \frac{\log(n+1)}{\log n}\right)$$

$$= \big(1 + o(1)\big)Z_{n^c,k}(t)$$

$$= \big(1 + o(1)\big)(1 - t)^k.$$

This completes the proof.   □

3. Nodes of small degree in the first level

Looking at the picture Theorem 2.1 shows us on the degree distribution one can naturally ask how many points of fixed degree remain in the lower levels at all. In this respect the first level and the other ones behave differently. It is easy to see that degree 1 nodes in level 1 correspond to the fixed points of the random permutation described in the Introduction. Hence their number has a Poisson limit distribution with parameter 1 without any normalization. More generally, let

$$X[n, d] = \big|\{i \in L_n(1) : \deg_n(i) = d\}\big|;$$

this is the number of nodes with degree d in the first level after n steps.

The main result of this section is the following limit theorem.

Theorem 3.1. $X[n, 1]$, $X[n, 2]$, ... *are asymptotically i.i.d. Poisson with mean* 1, *as* $n \to \infty$.

Proof. We will apply the method of moments in the following form.

For any real number a and nonnegative integer k let us define $(a)_0 = 1$, and $(a)_k = a(a - 1) \cdots (a - k + 1)$, $k = 1, 2, \ldots$. In order to verify the limiting joint distribution in Theorem 3.1 it suffices to show that

$$(1) \qquad \lim_{n \to \infty} \mathbb{E}\left(\prod_{i=1}^{d} \big(X[n, i]\big)_{k_i}\right) = 1$$

holds for every $d = 1, 2, \ldots$, and nonnegative integers k_1, ..., k_d. This can easily be seen from the following expansion of the joint probability generating function of the random variables $Y[n, 1]$, ..., $Y[n, d]$.

$$\mathbb{E}\left(\prod_{i=1}^{d} z_i^{Y[n,i]}\right) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_d=0}^{\infty} \mathbb{E}\left(\prod_{i=1}^{d} \big(X[n, i]\big)_{k_i}\right) \prod_{i=1}^{d} \frac{(z_i - 1)^{k_i}}{k_i!}.$$

In the proof we shall rely on the following obvious identities.

$$(2) \qquad (a + 1)_k - (a)_k = k\,(a)_{k-1},$$

$$(3) \qquad a\big[(a - 1)_k(b + 1)_\ell - (a)_k(b)_\ell\big] = \ell\,(a)_{k+1}(b)_{\ell-1} - k\,(a)_k(b)_\ell,$$

$$(4) \qquad \sum_{a=k}^{n} (a)_k = \frac{1}{k+1}\,(n + 1)_{k+1}.$$

Let us start from the conditional expectation of the quantity under consideration with respect to the sigma-field generated by the past of the process.

$$(5) \qquad \mathbb{E}\left(\prod_{i=1}^{d} \big(X[n+1, i]\big)_{k_i} \,\middle|\, \mathcal{F}_n\right) = \prod_{i=1}^{d} \big(X[n, i]\big)_{k_i} + \sum_{j=0}^{d} S_j,$$

where in the rightmost sum j equals 0, 1, ..., d, according that the new vertex at step $n + 1$ is connected to the root ($j = 0$), or to a degree j node in level 1. This happens with (conditional) probability $\dfrac{1}{n}$, $\dfrac{X[n, 1]}{n}$, ..., $\dfrac{X[n, d]}{n}$, respectively. That is,

$$S_0 = \frac{1}{n} \prod_{i=2}^{d} \big(X[n, i]\big)_{k_i} \left[\big(X[n, 1] + 1\big)_{k_1} - \big(X[n, 1]\big)_{k_1}\right],$$

and for $1 \leq j \leq d-1$

$$S_j = \frac{X[n,j]}{n} \prod_{i \neq \{j,j+1\}} \left(X[n,i]\right)_{k_i} \times$$
$$\times \left[\left(X[n,j]-1\right)_{k_j}\left(X[n,j+1]+1\right)_{k_{j+1}} - \left(X[n,j]\right)_{k_j}\left(X[n,j+1]\right)_{k_{j+1}}\right].$$

Finally,

$$S_d = \frac{X[n,d]}{n} \prod_{i=1}^{d-1}\left(X[n,i]\right)_{k_i}\left[\left(X[n,d]-1\right)_{k_d} - \left(X[n,d]\right)_{k_d}\right].$$

Let us apply (2) to S_0 with $k = k_1$, (3) to S_j with $k = k_j$, $\ell = k_{j+1}$ ($1 \leq j \leq d-1$), and (3) to S_d with $k = k_d$, $\ell = 0$, to obtain

(6)
$$S_0 = \frac{k_1}{n} \prod_{i=2}^{d}\left(X[n,i]\right)_{k_i}\left(X[n,1]\right)_{k_1-1},$$

$$S_j = \frac{k_{j+1}}{n} \prod_{i \neq \{j,j+1\}}\left(X[n,i]\right)_{k_i}\left(X[n,j]\right)_{k_j+1}\left(X[n,j+1]\right)_{k_{j+1}-1}$$

(7)
$$-\frac{k_j}{n} \prod_{i=1}^{d}\left(X[n,i]\right)_{k_i},$$

(8)
$$S_d = -\frac{k_d}{n} \prod_{i=1}^{d}\left(X[n,i]\right)_{k_i}.$$

In (6)–(7) it can happen that some of the k_j are zero, and, though $(a)_{-1}$ has not been defined, it always gets a zero multiplier, thus the expressions do have sense. Let us plug (6)–(8) into (5).

$$\mathbb{E}\left(\prod_{i=1}^{d}\left(X[n+1,i]\right)_{k_i} \,\middle|\, \mathcal{F}_n\right)$$
$$= \prod_{i=1}^{d}\left(X[n,i]\right)_{k_i}\left(1 - \frac{1}{n}\sum_{j=1}^{d}k_j\right) + \frac{k_1}{n}\prod_{i=2}^{d}\left(X[n,i]\right)_{k_i}\left(X[n,1]\right)_{k_1-1}$$
$$+ \sum_{j=1}^{d-1}\frac{k_{j+1}}{n}\prod_{i \neq \{j,j+1\}}\left(X[n,i]\right)_{k_i}\left(X[n,j]\right)_{k_j+1}\left(X[n,j+1]\right)_{k_{j+1}-1}.$$

Introducing

$$E(n,k_1,\ldots,k_d) = \mathbb{E}\left(\prod_{i=1}^{d}\left(X[n,i]\right)_{k_i}\right), \quad K = k_1 + \cdots + k_d,$$

we have the following recursion.

$$E(n+1,k_1,\ldots,k_d) = \left(1 - \frac{K}{n}\right)E(n,k_1,\ldots,k_d) + \frac{k_1}{n}E(n,k_1-1,k_2,\ldots,k_d)$$
$$+ \sum_{j=1}^{d-1}\frac{k_{j+1}}{n}E(n,k_1,\ldots,k_j+1,k_{j+1}-1,\ldots,k_d),$$

or equivalently,

$$(9) \quad (n)_K E(n+1, k_1, \ldots, k_d) = (n-1)_K E(n, k_1, \ldots, k_d)$$

$$+ (n-1)_{K-1} \sum_{j=1}^{d} k_j E(k_1, \ldots, k_{j-1}+1, k_j-1, \ldots, k_d).$$

Based on (9), the proof can be completed by induction on the exponent vectors $(k_1, \ldots, k_n)$. We say that $\underline{k} = (k_1, k_2, \ldots, k_d)$ is majorized by $\underline{\ell} = (\ell_1, \ell_2, \ldots, \ell_d)$, if $k_d \leq \ell_d$, $k_{d-1} + k_d \leq \ell_{d-1} + \ell_d$, ..., $k_1 + \cdots + k_d \leq \ell_1 + \cdots + \ell_d$. This is a total order on $\mathbb{N}^d$.

Now, (1) is clearly holds for $\underline{k} = (1, 0, \ldots, 0)$, since $\mathbb{E} X[n, 1] = 1$ for every $n = 1, 2, \ldots$, which is obvious considering the fixed points of a random permutation.

In every term of the sum on the right hand side of (9) the argument of $E(\cdot)$ is majorized by $\underline{k} = (k_1, \ldots, k_d)$, hence the induction hypothesis can be applied to them. We get that

$$(n)_K E(n+1, \underline{k}) = (n-1)_K E(n, \underline{k}) + (n-1)_{K-1} K \big(1 + o(1)\big),$$

from which (4) gives that $(n-1)_K E(n, \underline{k}) \sim (n-1)_K$, that is,

$$\lim_{n \to \infty} E(n, k_1, \ldots, k_d) = 1,$$

as needed. $\qquad \square$

Turning to higher levels one finds the situation changed. Fixing a degree d we find, roughly speaking, that each node in level $k-1$ has a Poisson number of degree d children in level k (a freshly added node is considered the child of the old node it is connected to). Now, strong-law-of-large-numbers-type heuristics imply that the number of nodes with degree d in level $k \geq 2$ is approximately equal to $|L_n(k-1)|$, that is, their proportion is

$$\approx \frac{|L_n(k-1)|}{|L_n(k)|} \sim \frac{1}{k \log n} \, .$$

Another interesting problem worth of dealing with is the number of nodes with unusually high degree. In every fixed level Theorem 2.1 implies that the proportion of nodes with degree higher than $\log n$ is asymptotically negligible, but they must exist, since the maximal degree is approximately $\log_2 n = \log_2 e \log n$. How many of them are there? We are planning to return to this issue in a separate paper.

References

[1] **Alon, N. and J.H. Spencer,** *The Probabilistic Method, 2nd ed.*, Wiley, New York, 2000.
[2] **Backhausz, Á. and T.F. Móri,** Local degree distribution in scale free random graphs, *Electron. J. Probab.*, **16** (2011), 1465–1488.
http://www.math.washington.edu/~ejpecp/include/getdoc.php?id=
6160&article=2250&mode=pdf
[3] **Barabási, A.-L. and R. Albert,** Emergence of scaling in random networks, *Science*, **286** (1999), 509–512.
[4] **Devroye, L. and J. Lu,** The strong convergence of maximal degrees in uniform random recursive trees and dags, *Random Structures Algorithms*, **7** (1995), 1–14.
[5] **Drmota, M.,** *Random Trees*, Springer, Wien, 2009.
[6] **Fuchs, M., H.-K. Hwang and R. Neininger,** Profiles of random trees: Limit theorems for random recursive trees and binary search trees, *Algorithmica* **46** (2006), 367–407.
[7] **Katona, Zs.,** Levels of a scale-free tree, *Random Structures Algorithms*, **29** (2006), 194–207.
[8] **Móri, T.F.,** A surprising property of the Barabási-Albert random tree, *Studia Sci. Math. Hungar.*, **43** (2006), 263–271.
[9] **Smythe, R.T. and H.M. Mahmoud,** A survey of recursive trees, *Theory Probab. Math. Statist.*, **51** (1995), 1–27.

[10] **Szymański, J.,** On a nonuniform random recursive tree, *Ann. Discrete Math.*, **33** (1987), 297–306.

[11] **Tapia, M.A. and B.R. Myers,** Generation of concave node-weighted trees, *IEEE Trans. Circuit Theory*, **CT-14** (1967), 229–230.

DEPARTMENT OF PROBABILITY THEORY AND STATISTICS, FACULTY OF SCIENCE, EÖTVÖS LORÁND UNIVERSITY, PÁZMÁNY P. S. 1/C, H-1117 BUDAPEST, HUNGARY

E-mail address: `agnes@cs.elte.hu`

DEPARTMENT OF PROBABILITY THEORY AND STATISTICS, FACULTY OF SCIENCE, EÖTVÖS LORÁND UNIVERSITY, PÁZMÁNY P. S. 1/C, H-1117 BUDAPEST, HUNGARY

E-mail address: `moritamas@ludens.elte.hu`